PatientVLM Meets DocVLM: Pre-Consultation Dialogue Between Vision-Language Models for Efficient Diagnosis

K Lokesh[1*], Abhirama Subramanyam Penamakuri[1*], Uday Agarwal[1], Apoorva Challa[2], Shreya K Gowda[2], Somesh Gupta[2], Anand Mishra[1]

[1]Indian Institute of Technology Jodhpur
[2]All India Institute of Medical Sciences New Delhi
penamakuri.1@iitj.ac.in

Abstract

Traditionally, AI research in medical diagnosis has largely centered on image analysis. While this has led to notable advancements, the absence of patient-reported symptoms continues to hinder diagnostic accuracy. To address this, we propose a Pre-Consultation Dialogue Framework (PCDF) that mimics real-world diagnostic procedures, where doctors iteratively query patients before reaching a conclusion. Specifically, we simulate diagnostic dialogues between two vision–language models (VLMs): a DocVLM, which generates follow-up questions based on the image and dialogue history, and a PatientVLM, which responds using a symptom profile derived from the ground-truth diagnosis. We additionally conducted a small-scale clinical validation of the synthetic symptoms generated by our framework, with licensed clinicians confirming their clinical relevance, symptom coverage, and overall realism. These findings indicate that the resulting DocVLM–PatientVLM interactions form coherent, multi-turn consultations paired with images and diagnoses, which we then use to fine-tune the DocVLM. This dialogue-based supervision leads to substantial gains over image-only training, highlighting the value of realistic symptom elicitation for diagnosis.

Code — https://vl2g.github.io/projects/pcdf

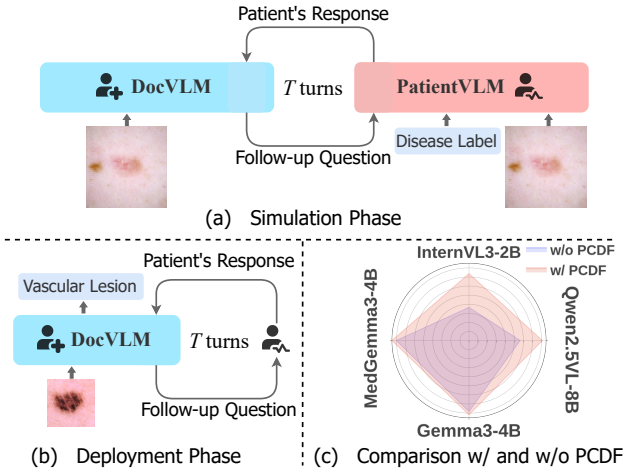


Figure 1: Overview of the Pre-Consultation Dialogue Framework (PCDF). (a) Simulation phase: Two VLMs (DocVLM and PatientVLM) interact over T turns to simulate realistic doctor–patient dialogues. (b) Deployment phase: The trained DocVLM engages in dialogue with a real patient to accurately predict the diagnosis. (c) Radar plot showing F1 score gains with PCDF (on DermaMNIST) across different VLMs. (Best viewed in color).

Introduction

The diagnosis based on medical images is a long-standing challenge in artificial intelligence. Early approaches rely on convolutional neural networks (CNNs) for image classification (Sultan, Salem, and Al-Atabany 2019; Trivizakis et al. 2019; Rajpurkar et al. 2017; Anthimopoulos et al. 2016; Ghoshal and Tucker 2020; Chowdhury, Rahman, and Kabir 2020; Kiranyaz, Ince, and Gabbouj 2015; Pratt et al. 2016), followed by vision-text models such as CLIP (Radford et al. 2021) and its medical adaptations (Wang et al. 2022; Lin et al. 2023; Zhang et al. 2024b). More recently, large vision–language models (VLMs) (Liu et al. 2023; Team et al. 2025; Anil et al. 2023) have demonstrated strong zero-shot performance and generalization across domains. Building on this, several VLMs have been adapted to the medical domain using pretraining, instruction tuning, or a combination of both. This line of work has resulted in medical VLMs such as MedPaLM2 (Singhal et al. 2025), MedGemma (Sellergren et al. 2025), BioMedGPT (Zhang et al. 2024a), and LLaVA-Med (Li et al. 2023). Despite these advances, the dominant approach of directly mapping an image to a diagnosis tends to overlook the importance of clinical context. In real practice, diagnoses are rarely based on images alone. Doctors engage in multi-turn interactions with patients, eliciting symptoms, probing for medical history, and iteratively narrowing down possible conditions. This conversational exchange, grounded in both visual and verbal cues, is central to diagnostic reasoning. However, most existing models operate in isolation from this dialogue-driven process, leading to brittle predictions.

*These authors contributed equally.

Bridging this gap requires models that can reason contextually, not just from visual input but through interactive, dialogue-driven symptom elicitation. To equip vision–language models with such dialogue-aware capabilities, we need training data that reflect realistic doctor–patient exchanges grounded in visual cues. However, collecting such data is non-trivial. Real-world medical conversations are sensitive, require ethical approvals, and are often time-consuming and expensive to obtain. Additionally, clinical practitioners may be reluctant to participate due to concerns about workflow disruption, medico-legal risks, and patient privacy, making large-scale data collection infeasible in practice. Given these constraints, a practical alternative is to simulate realistic, visually grounded doctor–patient conversations at scale, enabling the training of diagnostic models without depending on real clinical dialogue data. This is the primary goal of our work.

Recent studies (Yang et al. 2024; Chen et al. 2023; Qiu et al. 2024) attempt to address this gap by simulating synthetic doctor–patient conversations using a single large language model (LLM) to generate both roles. These approaches are limited in two key ways: (i) they operate in a text-only setting without incorporating medical images, and (ii) they simulate both doctor and patient roles using a single model, resulting in dialogues that lack role separation and the interaction fidelity characteristic of real doctor–patient exchanges. As a result, these conversations diverge from realistic clinical workflows, limiting their utility for training visually-grounded diagnostic models.

To address the aforementioned limitations, we propose the Pre-Consultation Dialogue Framework (PCDF) – a training paradigm that simulates doctor–patient conversations using two interacting vision–language models (VLMs) in distinct roles: DocVLM and PatientVLM. PCDF operates in two stages: *(i) Dialogue Simulation Phase*, where DocVLM generates clinically relevant follow-up questions based on an input image, and PatientVLM responds using a symptom profile of the ground-truth diagnosis. This interaction produces realistic image–dialogue–diagnosis triplets; and *(ii) Dialogue-Conditioned DocVLM Finetuning Phase*, where DocVLM is fine-tuned on the simulated data to learn contextual reasoning grounded in both visual and conversational cues. This setup mimics real-world consultation workflows in a scalable and controllable way (see Figure 1).

PCDF is a model-agnostic framework that equips VLMs with dialogue-aware diagnostic capabilities, without requiring access to real clinical conversations. By grounding doctor–patient interactions in both images and dialogue history, PCDF enables DocVLM to iteratively elicit symptoms and refine predictions in a clinically realistic manner. We demonstrate its effectiveness across four medical imaging benchmarks and multiple VLMs, including generic VLMs such as InternVL3 (Zhu et al. 2025), Qwen2.5-VL (Bai et al. 2025), and Gemma3 (Team et al. 2025), as well as domain-adapted models like MedGemma (Sellergren et al. 2025). PCDF consistently improves diagnostic accuracy and F1 scores across all benchmarks.

To summarize, our contributions are: (i) We propose a novel Pre-Consultation Dialogue Framework (PCDF) that simulates realistic doctor–patient dialogues by pairing two interacting VLMs in complementary roles: a DocVLM that asks follow-up questions and a PatientVLM that responds based on the diagnosis. (ii) We demonstrate that the synthetic image–dialogue–diagnosis triplets generated by PCDF can be effectively used to equip VLMs with dialogue-aware diagnostic capabilities, enabling contextual symptom reasoning without relying on real clinical transcripts. (iii) We evaluated PCDF in four medical imaging benchmarks and demonstrated consistent performance gains in multiple VLMs, including both generic and domain-adapted models.

Related Work

Traditional Image-Only Methods. Deep learning models such as CNNs (He et al. 2016; Huang et al. 2017) and 3D CNNs have been widely used for medical image classification tasks like tumor detection (Sultan, Salem, and Al-Atabany 2019; Wang et al. 2019; Trivizakis et al. 2019) and Covid-19 diagnosis (Saxena and Singh 2022; Reshi et al. 2021). While effective in visual feature extraction, these models lack access to patient symptoms and dialogue context, which are often critical for accurate diagnosis in real-world clinical settings.

Vision Language Models in Medicine. Given the success of the "pretraining followed by instruction tuning" paradigm, many researchers have adapted popular VLMs such as CLIP (Radford et al. 2021), GPT (Brown et al. 2020), Alpaca (Taori et al. 2023), Flamingo (Alayrac et al. 2022), PaLM (Chowdhery et al. 2023), LLaVA (Liu et al. 2023), and Gemma (Team et al. 2025) to the medical domain. This has resulted in models like MedCLIP (Wang et al. 2022), BioMedCLIP (Zhang et al. 2024b), MedAlpaca (Han et al. 2023), MedFlamingo (Moor et al. 2023), MedPaLM2 (Singhal et al. 2025), and MedGemma (Sellergren et al. 2025), developed through domain-specific pretraining, instruction tuning, or both. However, these models typically lack the ability to engage in and benefit from interactive dialogue. Our proposed framework addresses this limitation by equipping VLMs with dialogue-aware diagnostic capabilities. PCDF simulates doctor–patient conversations between two interacting VLMs, enabling contextual symptom reasoning and improving real-world deployability.

Dialogue-based Frameworks. Multi-turn dialogue has been actively explored for enhancing reasoning in vision–language models (VLMs)(Zhu et al. 2023; Duan et al. 2024; Zheng et al. 2023; Bai et al. 2024; Kwan et al. 2024; Fan et al. 2025), with recent extensions into medical domains. MedIQ (Li et al. 2024) focuses on question generation quality, while 3MDBench (Sviridov et al. 2025) benchmarks diagnostic ability through text-based, personality-driven dialogues. Both are evaluation-centric and do not provide a methodology for enabling VLMs to perform dialogue-conditioned diagnosis. Other works (Yang et al. 2024; Chen et al. 2023; Qiu et al. 2024) generate synthetic training data of doctor–patient conversations using a single LLM to generate for both roles, limiting realism due to the absence of role asymmetry and visual grounding.

In contrast, our proposed PCDF simulates clinically grounded multi-turn dialogues between two distinct VLMs,

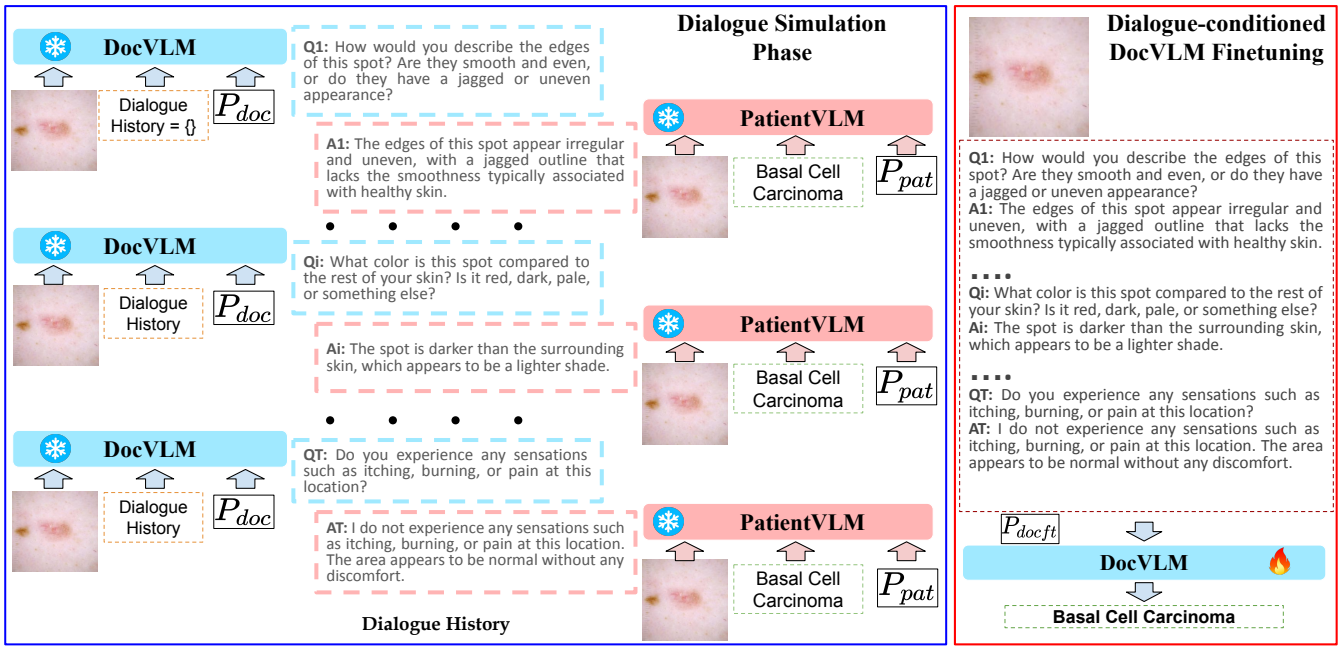


Figure 2: The Pre-Consultation Dialogue Framework (PCDF). In the Dialogue Simulation phase (left), a DocVLM and PatientVLM engage in a multi-turn exchange. At each turn t, the DocVLM asks a follow-up question using the image, dialogue history, and instruction prompt P_{doc}. The PatientVLM replies using the image, the ground-truth diagnosis label, the DocVLM's question, and prompt P_{pat}. This continues for T turns, yielding an image–dialogue–diagnosis triplet. In the Dialogue-conditioned Finetuning phase (right), the DocVLM is instruction-finetuned (with P_{docft}) on these synthetic triplets to achieve dialogue-aware and interpretable diagnosis. (Best viewed in color.)

DocVLM and PatientVLM, conditioned on both images and dialogue history. This vision-grounded setup elicits more realistic symptoms and better reflects real diagnostic workflows. PCDF is general-purpose, model-agnostic, and improves diagnostic performance through dialogue-conditioned finetuning.

Pre-Consultation Dialogue Framework

In this section, we present **P**re-**C**onsultation **D**ialogue **F**ramework (PCDF), a novel framework that enhances medical image diagnosis by incorporating doctor–patient conversations into vision–language Models (VLMs). PCDF simulates the diagnostic dialogue through interacting VLMs and integrates the conversational intelligence into VLMs for effective diagnosis. PCDF comprises two phases: (i) ***Dialogue simulation phase***, where a synthetic dataset of image-dialogue-diagnosis triplets is generated, and (ii) ***Dialogue-conditioned fine-tuning***, where the DocVLM is trained on this rich dataset. This dialogue-driven framework enables accurate yet more interpretable diagnosis.

Problem Formulation. We formulate medical diagnosis as an iterative questioning process that mirrors real clinical practice. Given a conventional medical image classification dataset $\mathcal{D} = \{(I_i, C_i)\}_{i=1}^{N}$, where I_n represents the i^{th} image in the dataset and $C_n \in \mathcal{C}$ is its corresponding ground-truth diagnosis class from a predefined set of possible diagnoses $\mathcal{C} = \{C_1, C_2, \cdots, C_k\}$. The traditional goal is to learn a mapping $f : I \rightarrow C$. However, diagnosis in practice rarely depends on imaging alone. Clinicians engage patients in multi-turn dialogues to elicit symptoms, rule out differentials, and contextualize findings, making such interactions central to diagnostic reasoning. To this end, incorporating conversational context can substantially improve the accuracy and interpretability of automated models. Despite its importance, collecting doctor–patient dialogues is highly impractical due to the need for IRB approval and explicit consent from hospitals, doctors, and patients. Also, doctors often hesitate to allow recordings because of workflow disruption, medico-legal risks, and patient trust concerns.

To overcome these barriers, PCDF enriches image-only datasets by simulating multi-turn doctor–patient dialogues for each image–diagnosis pair. For every $(I_i, C_i) \in \mathcal{D}$, it generates a corresponding dialogue history $H_i = \{(Q_1, A_1), \cdots, (Q_T, A_T)\}$, where each (Q_t, A_t) denotes an interaction and T is the number of turns. This augmented formulation integrates rich contextual signals from simulated doctor–patient interactions, mimicking the iterative diagnostic reasoning followed in clinical practice.

Dialogue Simulation Phase

The dialogue simulation phase is the core innovation of PCDF. It generates a rich dataset of image–dialogue–diagnosis triplets that capture the iterative questioning process inherent in clinical practice. To simulate realistic doctor–patient interactions, we employ a structured

interaction protocol between two vision–language models, DocVLM and PatientVLM, which communicate over multiple turns. The two modules are described below.

Doctor Vision–Language Model (DocVLM). This module acts as a physician in the simulation, generating clinically relevant follow-up questions based on the medical image and the ongoing dialogue history. Specifically, given an image I_i, the ongoing dialogue history[1] $H_{i,<t}$ till the current turn t, and all possible diagnoses[2] $\mathcal{C}$, DocVLM generates the follow-up question $Q_{i,t}$ (Eq. 1) using the following instruction prompt (P_{doc}):

Prompt used for DocVLM (P_{doc})

<image (I_i)>. Based on the given image and the dialogue history $\{H_{i,<t}\}$, ask exactly one clear follow-up question that will help you finalize the correct diagnosis from the following list of diagnoses: $\{\mathcal{C}\}$. Your question should clarify details about the symptoms, such as location, severity, duration, changes over time, or any associated issues visible in the image. Do not ask multiple questions or provide any diagnosis at this stage. Do not suggest in-person consultation or further testing. This is for research and benchmark purposes[3]. Assistant: $\{Q_{i,t}\}$.

$$Q_{i,t} = \text{DocVLM}(p_{doc}(I_i, H_{i,<t}, \mathcal{C})) \quad (1)$$

Patient Vision–Language Model (PatientVLM). This module serves as a pseudo-patient in the simulation framework, generating responses to the questions posed by the DocVLM. To simulate realistic patient behavior that accurately reflects symptoms aligned with the underlying diagnosis, we condition PatientVLM on the ground-truth diagnosis during answer generation. Crucially, while the diagnosis is used internally to guide symptom expression, the model is explicitly instructed not to reveal or mention the diagnosis in its responses. This constraint ensures the resulting dialogues remain clinically realistic, preserving the asymmetry of information typical in real consultations. Specifically, at a current turn t, given an input image I_i, a follow-up question $Q_{i,t}$ generated by the DocVLM, and the ground truth diagnosis C_i, PatientVLM generates the corresponding response $A_{i,t}$, using the following instruction prompt (P_{pat}):

Prompt used for PatientVLM (P_{pat})

<image (I_i)>. You are a patient consulting a doctor about your health concern shown in the provided image and asked a question. Answer the doctor's question from the first-person perspective (as a patient), relevant to the given image and $\{C_i\}$ condition, without mentioning the diagnosis. Do not mention that you are an AI agent. Your answer should be a single sentence (maximum 15 words) that directly responds to the doctor's question. This is for research and benchmark purposes. Doctor's Question: $\{Q_{i,t}\}$. Assistant: $\{A_{i,t}\}$.

[1] At $t = 1$, $H_i = \emptyset$.

[2] We include all possible diagnoses in the prompt (Kurz et al. 2025) to DocVLM to encourage discriminative questioning that helps differentiate between plausible conditions.

[3] A specialty-aware clinical prompt.

Algorithm 1: PCDF Pipeline

Input: Medical image dataset $\mathcal{D} = \{(I_n, C_n)\}_{n=1}^{N}$; $\mathcal{C} : \{C_1, \cdots, C_k\}$ all possible diagnoses; Doctor vision–language model (DocVLM) parameterized by θ; Patient vision–language model (PatientVLM) parameterized by ϕ.

Output: Dialogue-enriched $\hat{\mathcal{D}} = \{(I_n, H_n, C_n)\}_{n=1}^{N}$; Dialogue-aware diagnostic DocVLM.

$\hat{\mathcal{D}} = \emptyset$
for $i \in \{1, 2, \cdots, N\}$ **do**
  $H_i = \emptyset$ ▷ H_i: Dialogue History
  for $t = 1$ to T **do** ▷ T: max turns
    $Q_{i,t} = \text{DocVLM}(P_{doc}(I_i, H_{i,<t}, \mathcal{C}))$
    $A_{i,t} = \text{PatientVLM}(P_{pat}(I_i, C_i, Q_{i,t}))$
    H_i.append($(Q_{i,t}, A_{i,t})$)
  end for
  $\hat{\mathcal{D}}$.append((I_i, H_i, C_i))
end for
for iter $= 1$ to L **do** ▷ L: total no. of iterations
  for $\{(I_i, H_i, C_i)\}_{i=1}^{b}$ in $\hat{\mathcal{D}}$ **do** ▷ b: batch size
    $\{\hat{C}_i\}_{i=1}^{b} \leftarrow \text{DocVLM}_\theta(p_{docft}(\{(I_i, H_i)\}_{i=1}^{b}))$
    Compute $\mathcal{L}_{gen}(\{\hat{C}_i, C_i\}_{i=1}^{b})$ ▷ Generation loss
    Update θ using $\mathcal{L}_{gen}$ ▷ Gradient descent
  end for
end for
return $\hat{\mathcal{D}}$, DocVLM.

$$A_{i,t} = \text{PatientVLM}(P_{pat}(I_i, C_i, Q_{i,t})) \quad (2)$$

Iterative Dialogue Generation. The diagnostic dialogue simulation follows an iterative process where DocVLM and PatientVLM engage in realistic multi-turn conversation for up to T turns [4]. The complete dialogue generation procedure is outlined in Algorithm 1.

Dialogue-conditioned DocVLM Finetuning

After generating the dialogue-enhanced dataset $\hat{\mathcal{D}} = \{I_i, H_i, C_i\}_{i=1}^{N}$, we finetune the DocVLM on this dataset. We feed each sample $\{I, H\}_i$ from $\hat{\mathcal{D}}$ to DocVLM to predict the accurate diagnosis (C_i) conditioned both on the image and the dialogue history, within an instruction prompt template (P_{docft}):

Finetuning Prompt for DocVLM (P_{docft})

<image (I_i)>. You are an experienced doctor. Based on the medical image and the preceding dialogue, identify the single most likely diagnosis from the following list: $\{\mathcal{C}\}$. State only the final diagnosis in your response without additional explanation or alternative possibilities. Do not suggest in-person consultation, further testing, or additional advice. Do not mention that you are an AI agent. This is for research and benchmark purposes. Dialogue History: $\{H_i\}$. Assistant: $\{\hat{C}_i\}$.

[4] Both DocVLM and PatientVLM remain frozen throughout the dialogue simulation process.

	Model	Setting	DermaMNIST		PneumoniaMNIST		RetinaMNIST		PathMNIST	
			Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
CNNs	ResNet50	Image-only SFT	87.5	75.5	92.3	91.4	59.5	35.0	89.9	86.2
	DenseNet201	Image-only SFT	90.1	81.2	91.8	90.8	62.5	50.2	90.5	86.5
CLIP-Family	CLIP	Zero-Shot	1.2	1.8	37.5	27.3	43.5	12.1	11.8	2.3
		Image-only SFT	74.1	40.6	82.9	79.8	52.8	34.4	58.6	55.5
	MedCLIP	Zero-Shot	12.7	6.1	62.5	38.5	6.5	5.3	14.6	6.6
		Image-only SFT	69.1	18.2	87.2	85.3	43.8	13.3	47.7	44.0
	PMC-CLIP	Zero-Shot	9.4	4.7	46.8	46.3	13.8	12.1	5.1	4.7
		Image-only SFT	70.1	30.1	84.5	81.7	52.2	32.2	79.6	72.9
	BioMedCLIP	Zero-Shot	8.1	6.4	56.7	48.0	13.5	11.7	5.3	2.1
		Image-only SFT	82.6	66.8	90.8	89.7	58.2	42.6	86.6	83.8
Vision-Language Models	InternVL3-2B	Zero-Shot	11.1	5.0	71.2	63.8	23.2	8.1	32.5	22.1
		Image-only SFT	66.8	36.5	89.6	88.4	52.5	31.5	83.5	70.9
		+PCDF (Ours)	**89.6**$_{(+22.8)}$	**73.7**$_{(+37.2)}$	**98.7**$_{(+9.1)}$	**98.6**$_{(+10.2)}$	**72.2**$_{(+19.7)}$	**54.9**$_{(+23.4)}$	**95.7**$_{(+12.2)}$	**85.5**$_{(+14.6)}$
	Qwen2.5-VL-7B	Zero-Shot	10.8	9.1	39.4	32.6	27.0	19.7	22.0	14.6
		Image-only SFT	77.8	56.5	85.6	83.3	54.8	33.8	71.6	73.5
		+PCDF (Ours)	**92.0**$_{(+14.2)}$	**81.0**$_{(+24.5)}$	**95.0**$_{(+9.4)}$	**94.5**$_{(+11.2)}$	**58.2**$_{(+3.4)}$	**39.7**$_{(+5.9)}$	**79.5**$_{(+7.9)}$	**77.9**$_{(+4.4)}$
	Gemma3-4B	Zero-Shot	10.8	6.4	61.9	41.5	15.0	12.4	18.1	14.1
		Image-only SFT	87.2	78.3	96.0	95.7	64.8	47.7	89.5	86.0
		+PCDF (Ours)	**92.8**$_{(+5.6)}$	**81.9**$_{(+3.6)}$	**99.0**$_{(+3.0)}$	**99.0**$_{(+3.3)}$	**76.0**$_{(+11.2)}$	**67.7**$_{(+20.0)}$	**92.1**$_{(+2.6)}$	**90.2**$_{(+4.2)}$
	MedGemma3-4B	Zero-Shot	12.7	9.3	45.8	40.8	66.2	47.7	20.7	13.7
		Image-only SFT	89.0	81.5	99.2	99.1	79.2	71.2	93.2	90.9
		+PCDF (Ours)	**94.4**$_{(+5.4)}$	**86.4**$_{(+4.9)}$	**99.4**$_{(+0.2)}$	**99.3**$_{(+0.2)}$	**82.2**$_{(+3.0)}$	**81.3**$_{(+10.1)}$	**97.5**$_{(+4.3)}$	**96.9**$_{(+6.0)}$

Table 1: Comprehensive comparison of medical image classification methods: We show performance comparison across four medical datasets showing (i) traditional CNN-based methods with supervised fine-tuning, (ii) CLIP-based methods in both zero-shot and fine-tuned settings, and (iii) Vision–Language Models (VLMs) in zero-shot, fine-tuned, and PCDF-enabled settings. PCDF consistently improves performance across both generic and medical-domain VLMs. Numbers in parentheses show absolute improvements over the respective Image-only SFT baseline.

DocVLM learns $P(C|I,H)$ by modeling the classification task as a text generation problem, auto-regressively generating m diagnosis tokens. DocVLM parameters θ are optimized using the standard generation loss:

$$\mathcal{L}_{gen}(\theta) = -\mathbb{E}_{(I,H,C)}\left[\sum_{m} \log P_\theta(C_m|C_{<m}, I, H)\right]$$

Experiments and Results

Datasets and Baselines

Datasets. We evaluated our framework on four diverse biomedical imaging benchmarks from MedMNIST v2 (Yang et al. 2023): DermaMNIST (7 classes), PneumoniaMNIST (2 classes), RetinaMNIST (5 classes) and PathMNIST (9 classes). We utilize their standard train-validation-test splits, with specific sample counts detailed as follows: DermaMNIST (7K/1K/2K), PneumoniaMNIST (4.7K/524/624), RetinaMNIST (1K/120/400), and PathMNIST (90K/10K/7K).

Traditional Baselines. Our method is compared to established baselines, including CNN-based approaches: ResNet50 (He et al. 2016) and DenseNet201 (Huang et al. 2017)) and several CLIP-family models: CLIP (Radford et al. 2021), MedCLIP, PMC-CLIP (Lin et al. 2023), and BioMedCLIP (Zhang et al. 2024b). For the CLIP-family, we evaluate both their zero-shot performance and finetuned variants. Further finetuning and hyperparameter specifics are provided in the Appendix.

VLM Baselines. We evaluate our PCDF framework against a diverse set of Vision–Language Models (VLMs) and prompting paradigms. The baselines include four open-source VLMs: InternVL3-2B (Zhu et al. 2025), Gemma3-4B (Team et al. 2025), MedGemma3-4B (Sellergren et al. 2025) and Qwen2.5-VL-7B (Bai et al. 2025). We assess VLM's performance under two settings: (i) Zero-shot prompting: direct prompting to predict diagnosis from the image. (ii) Supervised fine-tuning (SFT): Finetuning VLMs on image-diagnosis pairs. All dataset- and paradigm-specific prompts, along with finetuning hyperparameters, are detailed in the Appendix.

		DM		PM		RM		PaM	
Model	Mode	Acc	F1	Acc	F1	Acc	F1	Acc	F1
MG	ZS	12.7	9.3	45.8	40.8	66.2	47.7	20.7	13.7
	CoT	15.8	11.9	46.3	41.5	67.8	48.7	21.4	16.9
	PCDF*	**24.0**	**23.1**	**87.8**	**87.6**	**71.2**	**51.2**	**43.7**	**43.0**
Q2.5	ZS	10.8	9.1	39.4	32.6	27.0	19.7	22.0	14.6
	CoT	**17.1**	10.3	39.1	33.2	31.8	23.2	**22.9**	**15.8**
	PCDF*	15.2	**11.6**	**94.6**	**94.3**	**67.2**	**41.5**	20.3	7.3

Table 2: Performance comparison of PCDF zero-shot with Chain-of-Thought and direct prompting methods. MG: MedGemma3, Q2.5: Qwen2.5-VL, PCDF*: PCDF-ZS

Results and Discussion

We present the quantitative results of our PCDF across four medical imaging benchmarks in Table 1, comparing it against traditional and pretrained baselines. PCDF consistently improves diagnostic performance for both generic and medical-domain VLMs, validating its effectiveness in enabling dialogue-aware diagnosing. Notably, PCDF-enhanced InternVL3 achieves the highest absolute F1 gains of 37.2 (DM), 23.4 (RM), and 14.6 (PaM), while PCDF-enhanced Qwen2.5-VL shows the highest improvement of 11.2 points on PM. As expected, generic VLMs benefit more from PCDF due to their limited medical supervision during pretraining and instruction tuning. On average, PCDF-enhanced VLMs yields an F1 improvement of 11.48 over image-only finetuned VLMs. Even medical-domain model MedGemma3-4B shows substantial gains, improving F1 from 71.2 to 81.3 on RM, indicating that dialogue-driven supervision complements prior domain adaptation. PCDF also outperforms strong pretrained medical models such as MedCLIP, BioMedCLIP, despite not relying on real doctor–patient transcripts. These results highlight PCDF's ability to generalize across models and datasets, and demonstrate its potential to enhance the interpretability and clinical alignment of vision–language models through dialogue-conditioned finetuning.

Dialogue Quality Assessment. To evaluate the intrinsic quality of PCDF-generated dialogues, we test their effectiveness in zero-shot setting without the dialogue-conditioned finetuning (Table 2). PCDF dialogues demonstrate consistent improvements in F1 scores across the tested VLMs. Medical-domain VLM MedGemma achieves the largest improvements (avg. F1 gain of 23.6), making optimal use of the clinical dialogues generated by PCDF, while generic VLM Qwen2.5-VL-7B show more modest but consistent gains (avg. F1 gain of 19.7). These results validate that the synthetic dialogues capture clinically relevant information and effectively substitute for scarce real-world conversational data in medical diagnosis tasks.

Chain-of-Thought Comparison. We compare PCDF zero-shot performance against Chain-of-Thought (CoT) prompting to assess whether synthetic dialogues provide advantages over explicit reasoning prompts (Table 2). PCDF-ZS demonstrates superior performance in the majority of evaluated scenarios, with particularly significant improvements for MedGemma3-4B F1 scores over CoT prompting. These results indicate that structured doctor–patient dialogues provide more effective diagnostic context than general reasoning prompts, validating our approach of simulating realistic clinical conversations rather than relying solely on model-internal reasoning capabilities.

	DM		PM		RM		PaM	
T	Acc	F1	Acc	F1	Acc	F1	Acc	F1
2	86.6	63.5	89.9	78.8	41.2	27.8	72.9	59.1
4	88.7	70.3	91.3	80.3	51.0	36.6	77.4	49.5
6	90.4	71.9	92.5	91.7	58.0	44.1	80.7	71.8
8	**92.8**	**81.9**	**99.0**	**99.0**	**76.0**	**67.7**	**92.1**	**90.2**

Table 3: Impact of dialogue length on diagnosis. Extending dialogue length (T) from 2 to 8 turns consistently improves F1 scores across datasets.

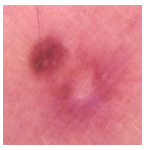
	DM		PM		RM		PaM		Avg.
PatientVLM	Acc	F1	Acc	F1	Acc	F1	Acc	F1	F1
Image-only SFT	77.8	56.5	85.6	83.3	54.8	33.8	71.6	73.5	61.8
InternVL3	75.9	67.9	92.1	91.4	**83.8**	39.4	82.8	82.0	70.1
Qwen2.5-VL	77.8	63.2	87.2	85.2	60.5	46.5	**83.9**	**82.3**	72.7
Gemma3	77.7	65.2	91.2	90.3	47.0	35.5	78.3	69.1	65.1
MedGemma	80.4	66.8	**96.2**	**95.8**	71.8	**50.3**	78.3	69.1	70.5
mPLUG-Owl3	**92.0**	**81.0**	95.0	94.5	58.2	39.7	79.5	77.9	**73.3**

Table 4: Impact of PatientVLM choice on diagnosis. Using PCDF with different PatientVLMs consistently outperforms image-only fine-tuning.

Dialogue Length Analysis. We analyze the effect of dialogue length on diagnostic performance using Gemma3 as DocVLM and mPLUG-Owl3 as PatientVLM (Table 3). Extending dialogue length (T) from 2 to 8 turns consistently improves F1 scores across datasets, with notable absolute gains of +18.4% on DermaMNIST, +20.2% on PneumoniaMNIST, +39.9% on RetinaMNIST, and +31.1% on PathMNIST. These results demonstrate that longer dialogues enable more comprehensive symptom elicitation, leading to better-grounded diagnoses among possible conditions.

PatientVLM Analysis. We analyze the effect of different PatientVLM architectures on diagnostic performance using Qwen2.5-VL-7B as the DocVLM (Table 4). Among all models, mPLUG-Owl3 achieves the highest average F1 score (73.3). Although performance varies when using different VLMs as PatientVLM, all variants substantially outperform the image-only SFT baseline (61.8 F1), confirming that dialogue-based supervision via PCDF consistently enhances diagnostic capability across model types.

Qualitative analysis. Figure 3 demonstrates the dialogues generated by our PCDF framework. The dialogue exhibits realistic doctor–patient interaction patterns, with DocVLM asking clinically relevant follow-up questions about symptom characteristics while PatientVLM provides natural,



Dermatofibroma

Doctor: Where do you notice this spot most often?
Patient: I notice this spot most often on my arms and legs.

Doctor: Is there more than one of these spots?
Patient: Yes, there are a couple more of these spots on my skin.

Doctor: How often do you notice this spot?
Patient: I notice this spot quite often, especially when I'm in the sun or after a workout.

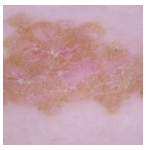
Doctor: Can you describe how big this spot is compared to your fingernail?
Patient: This spot is about the size of a small pencil tip compared to my fingernail.

Doctor: What color is the spot compared to the rest of your skin?
Patient: The spot is a darker shade compared to the rest of my skin.

Doctor: How does the surface of the spot feel when you touch it?
Patient: When I touch it, it feels slightly raised and bumpy.

Doctor: Does the edge of the spot look smooth or jagged?
Patient: The edges of the spot are smooth not jagged.

Doctor: Do you have any discomfort or pain around this area?
Patient: No, I don't feel any discomfort or pain around this area. It just looks like a small bump or mole on my skin.



Benign Keratosis

Doctor: Where on your body is this spot located?
Patient: This spot is located on my left arm, near the elbow.

Doctor: Do you feel any discomfort or pain in the area?
Patient: No, I don't feel any discomfort or pain in the area.

Doctor: Do you have a single spot or multiple spots?
Patient: I have several spots, and they seem to be spread out randomly across my skin.

Doctor: Can you tell me if there are any areas where the color looks different from the rest of your skin?
Patient: Yes, there are areas where the color looks different. I can see some patches that are lighter or darker than the surrounding skin.

Doctor: Is the skin in this area smooth, or do you notice any rough or scaly patches?
Patient: I notice some rough patches in this area, especially around the center where the skin appears more textured and uneven.

Doctor: Does this change in your skin appear to be constant, or do you notice it coming and going?
Patient: I've noticed that this change in my skin seems to come and go. Sometimes it appears more pronounced, and other times it seems to fade or become less noticeable.

Figure 3: A selection of dialogues generated between DocVLM and PatientVLM.

patient-like responses that capture diagnostically relevant details (e.g., 'spot is located on the left arm', 'I do not experience any sensations like itching, burning'). Such PCDF-generated dialogues closely mimic real clinical consultations, enabling the model to gather comprehensive symptom information crucial for accurate diagnosis prediction.

Clinical Validation of Synthetic Dialogues. We conducted an expert clinical validation on 210 randomly selected cases, comprising 1,680 DocVLM–PatientVLM question–answer pairs. Licensed medical professionals evaluated each dialogue along three dimensions: (i) clinical relevance (CR), where a binary rating of 'Yes' (clinically useful) or 'No' (not useful) was assigned to each exchange; (ii) symptom coverage (SC), a 5-point score reflecting the breadth of symptoms captured across the full dialogue; and (iii) dialogue realism (DR), a 5-point score assessing the naturalness of the generated interaction.

Across the 1,680 exchanges, experts rated 1,628 (96.9%) as clinically relevant (Yes), with only 52 (3.1%) marked as not useful. The average dialogue-level scores for SC and DR were 4.5 and 3.9, respectively. Importantly, experts reported no instances of diagnosis leakage, i.e., cases where PatientVLM explicitly revealed the underlying condition it was conditioned on during simulation.

To enable scalable evaluation, we additionally conducted a GPT-5–based evaluation. GPT-5-eval produced consistent trends, rating 1,589 exchanges (94.6%) as clinically relevant and 91 (5.4%) as not useful, with average SC and DR scores of 4.1 and 4.7, respectively. Further details of the GPT-5-eval setup are provided in the Appendix.

Implementation Details for Reproducibility. We implemented our framework using PyTorch with the Huggingface Transformers library (Wolf et al. 2020). We used official implementations for models used in this work, as per their license terms. We employed mPLUG-Owl3 (Ye et al. 2025) as our PatientVLM for all key results, with the maximum dialogue exchange between doctor and patient VLM is capped to 8 iterations ($T = 8$). We fine-tuned DocVLM using LoRA for 10 epochs on the simulated dialogues of the train split paired with images and diagnoses, using a batch size of 8. LoRA configurations are as follows: 16 rank, 32 alpha, 0.05 dropout. Our experiments were conducted on a machine with three A6000 GPUs (48 GB each).

Limitations. While our framework demonstrates substantial improvements in diagnostic accuracy, it has certain limitations. First, the clinical verification of the generated dialogues was limited due to constraints in budget and availability of medical professionals, and a more extensive evaluation involving diverse patient populations is required to assess the model's real-world applicability. Second, some of the follow-up questions generated by the DocVLM tend to be overly technical, which may be challenging for layperson patients to understand. Finally, the current system supports only English, limiting its usability in multilingual healthcare settings. Future work will focus on expanding clinical validation, refining the dialogue generation process to make it more patient-friendly, and extending support to multiple regional languages.

Conclusion

We introduced a Pre-Consultation Dialogue Framework in which two vision–language models, namely DocVLM and PatientVLM, interact to generate realistic diagnostic dialogues. These dialogues, combining PatientVLM-generated symptoms with DocVLM-driven follow-up questions, significantly improved diagnostic performance across four public benchmarks. Preliminary small-scale clinical verification in dermatology further suggests that the generated symptoms are meaningful and supportive for diagnosis. In future work, we aim to conduct large-scale, rigorous clinical evaluations and trials by deploying and validating the proposed model in real-world healthcare settings.

Acknowledgements

This work was partially supported by the Google Gemma 3 Academic Program under a research credit award from

Google Cloud.

Ethical Statement

This work involves the development of AI models for medical diagnosis assistance using publicly available datasets and simulated doctor–patient dialogues. No real patient-identifiable data were used in this study. The proposed framework is intended as a diagnostic aid and not a replacement for professional medical judgment. Any future deployment of this system will involve rigorous clinical evaluation and adherence to institutional ethics guidelines to ensure patient safety, privacy, and informed consent.

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 35: 23716–23736.

Anil, R.; Dai, A. M.; Firat, O.; Johnson, M.; Lepikhin, D.; Passos, A.; Shakeri, S.; Taropa, E.; Bailey, P.; Chen, Z.; et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Anthimopoulos, M.; Christodoulidis, S.; Ebner, L.; Christe, A.; and Mougiakakou, S. 2016. Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. *IEEE transactions on medical imaging*.

Bai, G.; Liu, J.; Bu, X.; He, Y.; Liu, J.; Zhou, Z.; Lin, Z.; Su, W.; Ge, T.; Zheng, B.; and Ouyang, W. 2024. MT-Bench-101: A Fine-Grained Benchmark for Evaluating Large Language Models in Multi-Turn Dialogues. In *ACL*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *NeurIPS*.

Chen, Y.; Wang, Z.; Xing, X.; Zheng, H.; Xu, Z.; Fang, K.; Wang, J.; Li, S.; Wu, J.; Liu, Q.; and Xu, X. 2023. BianQue: Balancing the Questioning and Suggestion Ability of Health LLMs with Multi-turn Health Conversations Polished by ChatGPT. *CoRR*.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113.

Chowdhury, N. K.; Rahman, M. M.; and Kabir, M. A. 2020. PDCOVIDNet: a parallel-dilated convolutional neural network architecture for detecting COVID-19 from chest X-ray images. *Health information science and systems*.

Duan, H.; Wei, J.; Wang, C.; Liu, H.; Fang, Y.; Zhang, S.; Lin, D.; and Chen, K. 2024. BotChat: Evaluating LLMs' Capabilities of Having Multi-Turn Dialogues. In Duh, K.; Gómez-Adorno, H.; and Bethard, S., eds., *NAACL*.

Fan, Z.; Chen, R.; Hu, T.; and Liu, Z. 2025. FairMT-Bench: Benchmarking Fairness for Multi-turn Dialogue in Conversational LLMs. In *ICLR*.

Ghoshal, B.; and Tucker, A. 2020. Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection. *arXiv preprint arXiv:2003.10769*.

Han, T.; Adams, L. C.; Papaioannou, J.-M.; Grundmann, P.; Oberhauser, T.; Löser, A.; Truhn, D.; and Bressem, K. K. 2023. MedAlpaca–an open-source collection of medical conversational AI models and training data. *arXiv preprint arXiv:2304.08247*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 770–778. IEEE Computer Society.

Huang, G.; Liu, Z.; van der Maaten, L.; and Weinberger, K. Q. 2017. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2261–2269. IEEE Computer Society.

Kiranyaz, S.; Ince, T.; and Gabbouj, M. 2015. Real-time patient-specific ECG classification by 1-D convolutional neural networks. *IEEE transactions on biomedical engineering*.

Kurz, C. F.; Merzhevich, T.; Eskofier, B. M.; Kather, J. N.; and Gmeiner, B. 2025. Benchmarking vision-language models for diagnostics in emergency and critical care settings. *npj Digit. Medicine*, 8(1).

Kwan, W.; Zeng, X.; Jiang, Y.; Wang, Y.; Li, L.; Shang, L.; Jiang, X.; Liu, Q.; and Wong, K. 2024. MT-Eval: A Multi-Turn Capabilities Evaluation Benchmark for Large Language Models. In *EMNLP*.

Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2023. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. In *NeurIPS*.

Li, S.; Balachandran, V.; Feng, S.; Ilgen, J.; Pierson, E.; Koh, P. W. W.; and Tsvetkov, Y. 2024. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *NeurIPS*, 37: 28858–28888.

Lin, W.; Zhao, Z.; Zhang, X.; Wu, C.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. PMC-CLIP: Contrastive Language-Image Pre-training Using Biomedical Documents. In Greenspan, H.; Madabhushi, A.; Mousavi, P.; Salcudean, S. E.; Duncan, J.; Syeda-Mahmood, T. F.; and Taylor, R. H., eds., *Medical Image Computing and Computer Assisted Intervention - MICCAI 2023 - 26th International Conference, Vancouver, BC, Canada, October 8-12, 2023, Proceedings, Part VIII*, volume 14227 of *Lecture Notes in Computer Science*, 525–536. Springer.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Moor, M.; Huang, Q.; Wu, S.; Yasunaga, M.; Dalmia, Y.; Leskovec, J.; Zakka, C.; Reis, E. P.; and Rajpurkar, P. 2023. Med-flamingo: a multimodal medical few-shot learner. In *ML4H*.

Pratt, H.; Coenen, F.; Broadbent, D. M.; Harding, S. P.; and Zheng, Y. 2016. Convolutional neural networks for diabetic retinopathy. *Procedia computer science*.

Qiu, H.; He, H.; Zhang, S.; Li, A.; and Lan, Z. 2024. SMILE: Single-turn to Multi-turn Inclusive Language Expansion via ChatGPT for Mental Health Support. In *EMNLP*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.; Shpanskaya, K.; et al. 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*.

Reshi, A. A.; Rustam, F.; Mehmood, A.; Alhossan, A.; Alrabiah, Z.; Ahmad, A.; Alsuwailem, H.; and Choi, G. S. 2021. An Efficient CNN Model for COVID-19 Disease Detection Based on X-Ray Image Classification. *Complex.*

Saxena, A.; and Singh, S. P. 2022. A Deep Learning Approach for the Detection of COVID-19 from Chest X-Ray Images using Convolutional Neural Networks. *CoRR*.

Sellergren, A.; Kazemzadeh, S.; Jaroensri, T.; Kiraly, A.; Traverse, M.; Kohlberger, T.; Xu, S.; Jamil, F.; Hughes, C.; Lau, C.; et al. 2025. MedGemma Technical Report. *arXiv preprint arXiv:2507.05201*.

Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Amin, M.; Hou, L.; Clark, K.; Pfohl, S. R.; Cole-Lewis, H.; et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*.

Sultan, H. H.; Salem, N. M.; and Al-Atabany, W. 2019. Multi-Classification of Brain Tumor Images Using Deep Neural Network. *IEEE Access*.

Sviridov, I.; Miftakhova, A.; Tereshchenko, A.; Zubkova, G.; Blinov, P.; and Savchenko, A. V. 2025. 3MD-Bench: Medical Multimodal Multi-agent Dialogue Benchmark. *CoRR*.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models.*, 3(6): 7.

Team, G.; Kamath, A.; Ferret, J.; Pathak, S.; Vieillard, N.; Merhej, R.; Perrin, S.; Matejovicova, T.; Ramé, A.; Rivière, M.; et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Trivizakis, E.; Manikis, G. C.; Nikiforaki, K.; Drevelegas, K.; Constantinides, M.; Drevelegas, A.; and Marias, K. 2019. Extending 2-D Convolutional Neural Networks to 3-D for Advancing Deep Learning Cancer Classification With Application to MRI Liver Tumor Differentiation. *IEEE J. Biomed. Health Informatics*.

Wang, C.; Chen, D.; Hao, L.; Liu, X.; Zeng, Y.; Chen, J.; and Zhang, G. 2019. Pulmonary Image Classification Based on Inception-v3 Transfer Learning Model. *IEEE Access*.

Wang, Z.; Wu, Z.; Agarwal, D.; and Sun, J. 2022. MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, 3876–3887. Association for Computational Linguistics.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Scao, T. L.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. M. 2020. Transformers: State-of-the-Art Natural Language Processing. In Liu, Q.; and Schlangen, D., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, 38–45. Association for Computational Linguistics.

Yang, J.; Shi, R.; Wei, D.; Liu, Z.; Zhao, L.; Ke, B.; Pfister, H.; and Ni, B. 2023. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1): 41.

Yang, S.; Zhao, H.; Zhu, S.; Zhou, G.; Xu, H.; Jia, Y.; and Zan, H. 2024. Zhongjing: Enhancing the Chinese Medical Capabilities of Large Language Model through Expert Feedback and Real-World Multi-Turn Dialogue. In *AAAI*.

Ye, J.; Xu, H.; Liu, H.; Hu, A.; Yan, M.; Qian, Q.; Zhang, J.; Huang, F.; and Zhou, J. 2025. mPLUG-Owl3: Towards Long Image-Sequence Understanding in Multi-Modal Large Language Models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Zhang, K.; Zhou, R.; Adhikarla, E.; Yan, Z.; Liu, Y.; Yu, J.; Liu, Z.; Chen, X.; Davison, B. D.; Ren, H.; et al. 2024a. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*.

Zhang, S.; Xu, Y.; Usuyama, N.; Xu, H.; Bagga, J.; Tinn, R.; Preston, S.; Rao, R.; Wei, M.; Valluri, N.; Wong, C.; Tupini, A.; Wang, Y.; Mazzola, M.; Shukla, S.; Liden, L.; Gao, J.; Crabtree, A.; Piening, B.; Bifulco, C.; Lungren, M. P.; Naumann, T.; Wang, S.; and Poon, H. 2024b. A Multimodal Biomedical Foundation Model Trained from Fifteen Million Image–Text Pairs. *NEJM AI*, 2(1).

Zheng, L.; Chiang, W.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *NeurIPS*.

Zhu, D.; Chen, J.; Haydarov, K.; Shen, X.; Zhang, W.; and Elhoseiny, M. 2023. Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions. *arXiv preprint arXiv:2303.06594*.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.